

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Data Augmentation with Diversified Rephrasing for Low-Resource Neural Machine Translation

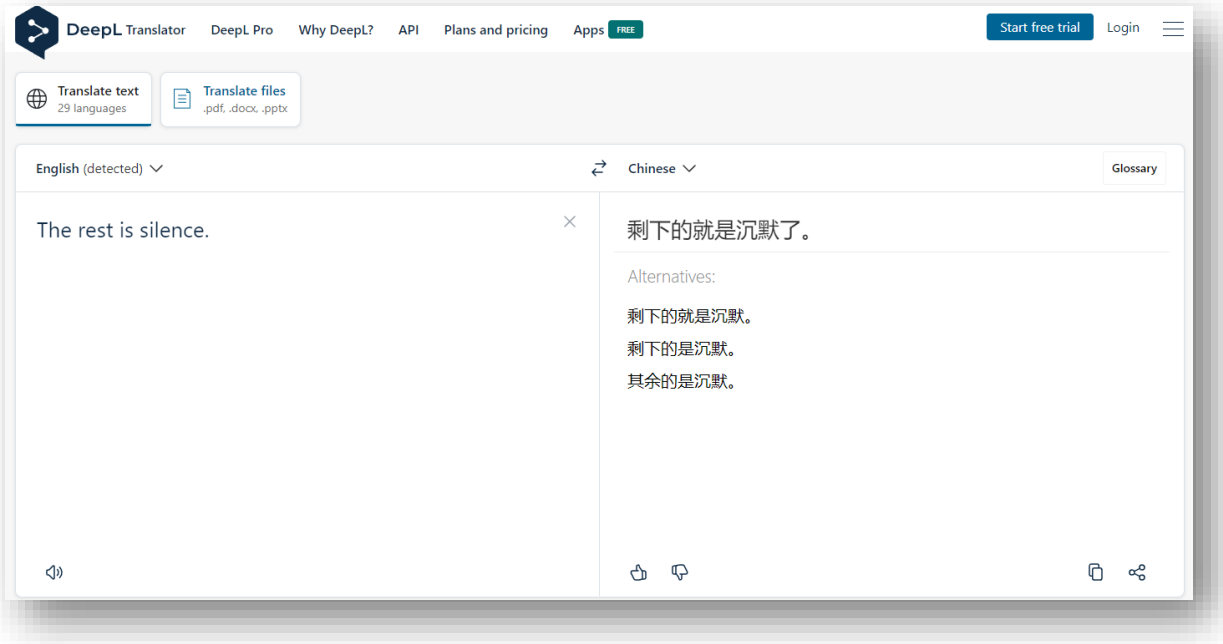
Yuan Gao, Feng Hou, Huia Jahnke, Ruili Wang*

School of Mathematical and Computational Sciences
Massey University
Auckland, New Zealand
4 September 2023

* Corresponding author

Introduction

- Machine Translation (MT)
- Neural Machine Translation (NMT)
- Low-resource
- Data Augmentation



Related Work



- ❖ Token-level Methods
  - Perform token-level manipulations
  - May change the original meanings and lack syntactic diversity
- ❖ Sentence-level Methods
  - Take natural sentences as input and generate synthetic corresponding sentences using pre-trained models
  - The augmented data are usually of low quality

Objective

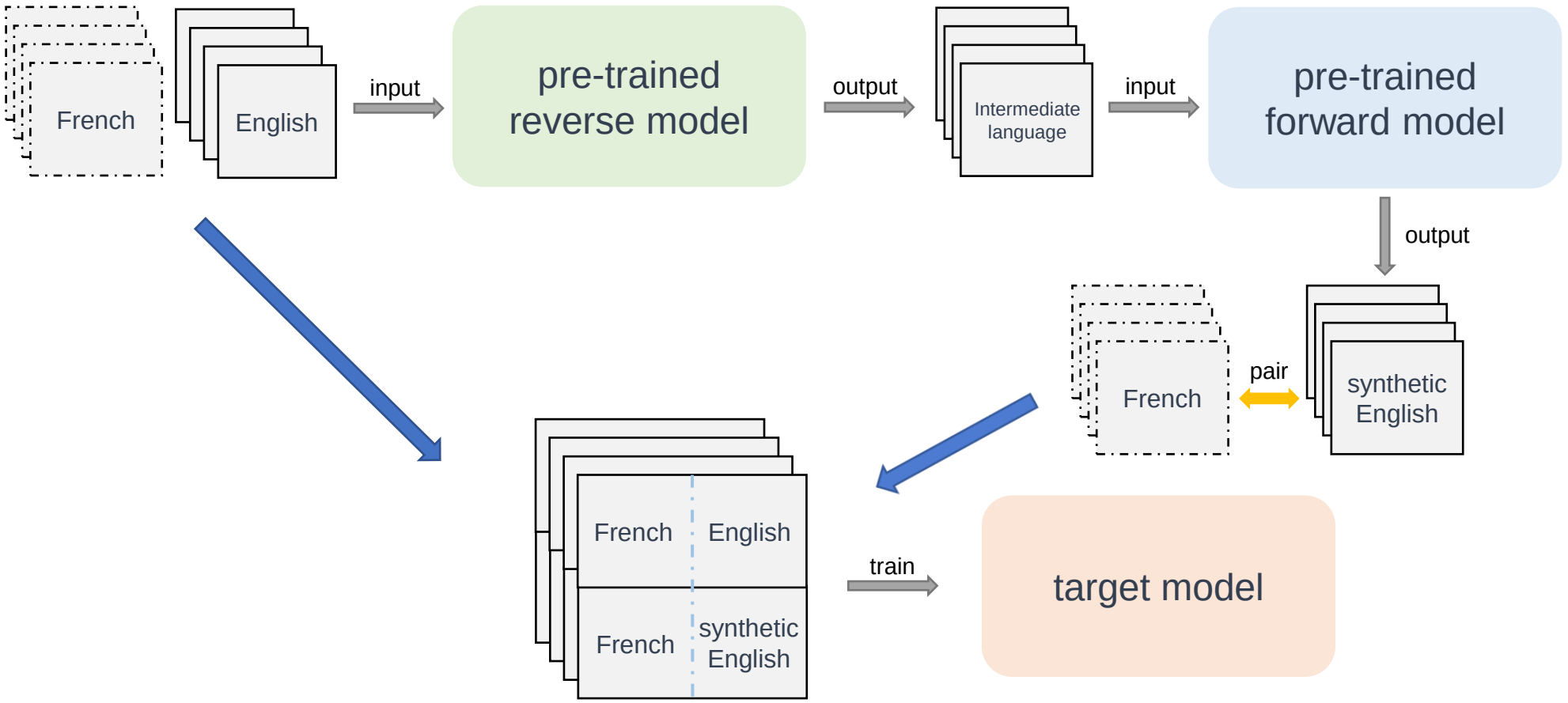


- Generate high-quality and meaning-matched parallel training data

Solution

- Introduce pre-trained high-quality translation models to rephrase the existing data

Proposed Method



Bidirectional **T**ranslation-based **D**ata **A**ugmentation (BiTDA)

Datasets

- IWSLT 2014 Hebrew → English
- IWSLT 2015 Vietnamese → English
- WMT 2016 Romanian → English
- WMT 2016 Russian → English
- WMT 2017 Finnish → English
- WMT 2017 Latvian → English
- WMT 2018 Turkish → English
- Google Site Korean → English

Evaluation Metric

- SacreBLEU

Experimental Results

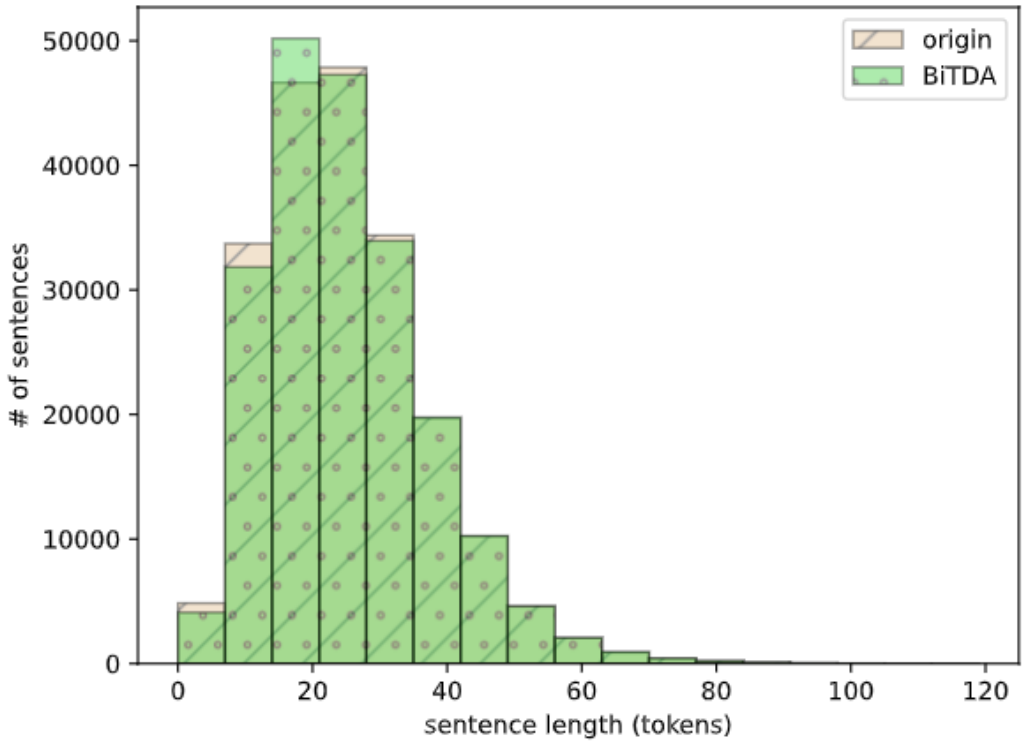


	Vi→En	He→En	Tr→En	Ro→En	Ru→En	Fi→En	Lv→En	Ko→En
Baseline	31.64	36.52	21.86	34.08	28.69	28.01	17.20	5.26
+ WordDropout	31.62	36.67	21.92	34.16	28.15	28.12	17.32	5.46
+ Swap	31.63	36.56	21.94	34.22	28.92	28.31	17.52	5.37
+ SwitchOut	32.35	36.93	22.28	33.86	28.13	28.33	17.10	5.00
+ BPEDropout	32.73	37.66	22.95	34.83	28.94	27.55	17.61	5.86
+ BiTDA-de (▲)	32.33	37.20	22.72	**35.07**	**30.01**	**28.57**	**17.75**	5.63
+ BiTDA-fr	32.37	37.23	22.63	34.75	28.95	27.24	16.95	5.54
+ BiTDA-(de+fr)	**32.96**	**37.72**	**23.56**	34.63	29.87	28.22	17.42	**5.89**
▲ + BPEDropout	**33.49**	**38.47**	23.40	**35.20**	29.98	**29.01**	**17.98**	**6.21**
▲ + MLS	33.19	37.38	22.90	34.38	29.64	28.74	17.82	5.02

Analysis



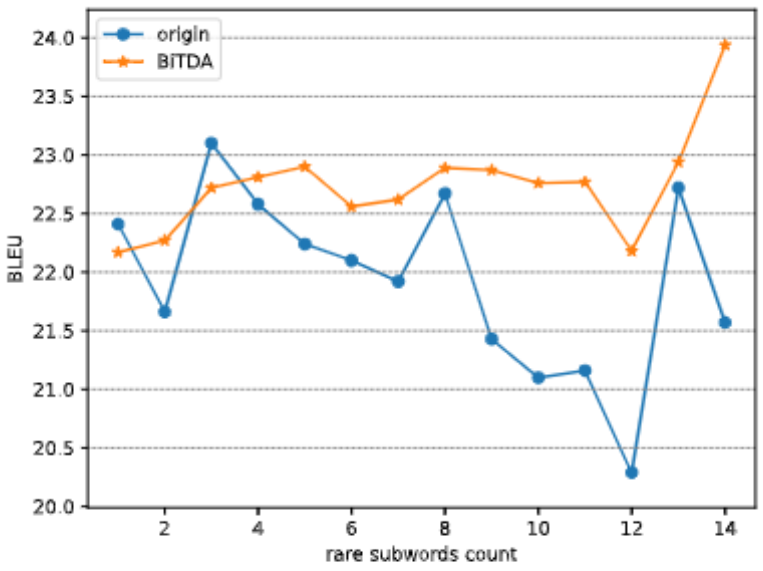
- Translationese



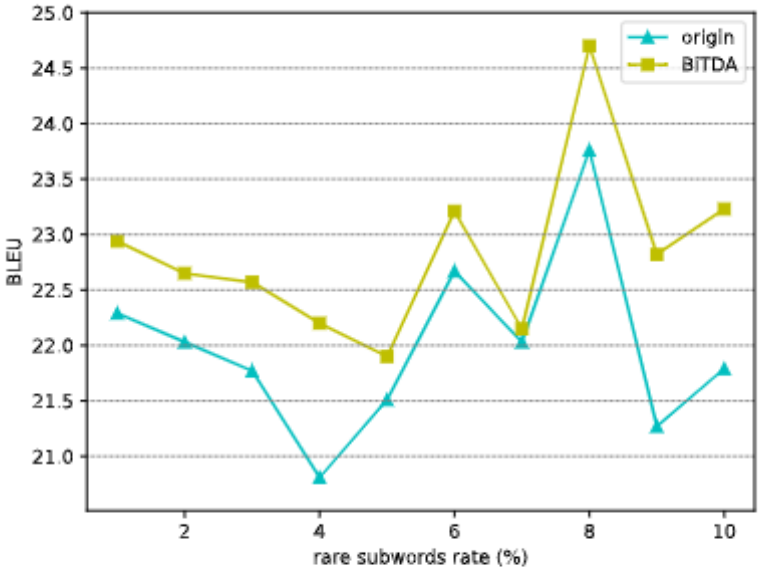
	Baseline	BiTDA-de	BiTDA-de	BiTDA-double
Tr-orig	17.17	17.96	17.80	18.51
En-orig	25.90	27.57	27.23	28.21

Analysis

- Rare Subwords



(a)



(b)

Analysis



❖ Case Study

Original:	Ten years ago, when a local bank launched its first credit card, only one shop in Bucharest's downtown was able to accept electronic payments.
BiTDA-de:	When a local bank introduced its first credit card ten years ago, only one shop in downtown Bucharest could accept electronic payments.
BiTDA-fr:	Ten years ago, when a local bank started its first credit card, a single store in Bucharest, in the center-city was able to accept electronic payments.
Original:	Some foresee a growth of up to 500 per cent by the end of the year for transactions originating in Romania.
BiTDA-de:	Some expect up to 500 percent growth in transactions originating in Romania by the end of the year.
BiTDA-fr:	Some are forecasting a growth rate of up to 500%, at the end of the year for transactions from Romania.

Conclusion



- We propose BiTDA, a simple yet effective data augmentation method for low-resource translation.
- To generate high-quality and meaning-matched parallel training samples, we introduce two pre-trained high-resource translation models.
- Experimental results validate the consistent effectiveness of our method across various low-resource benchmark datasets.

Thank you !